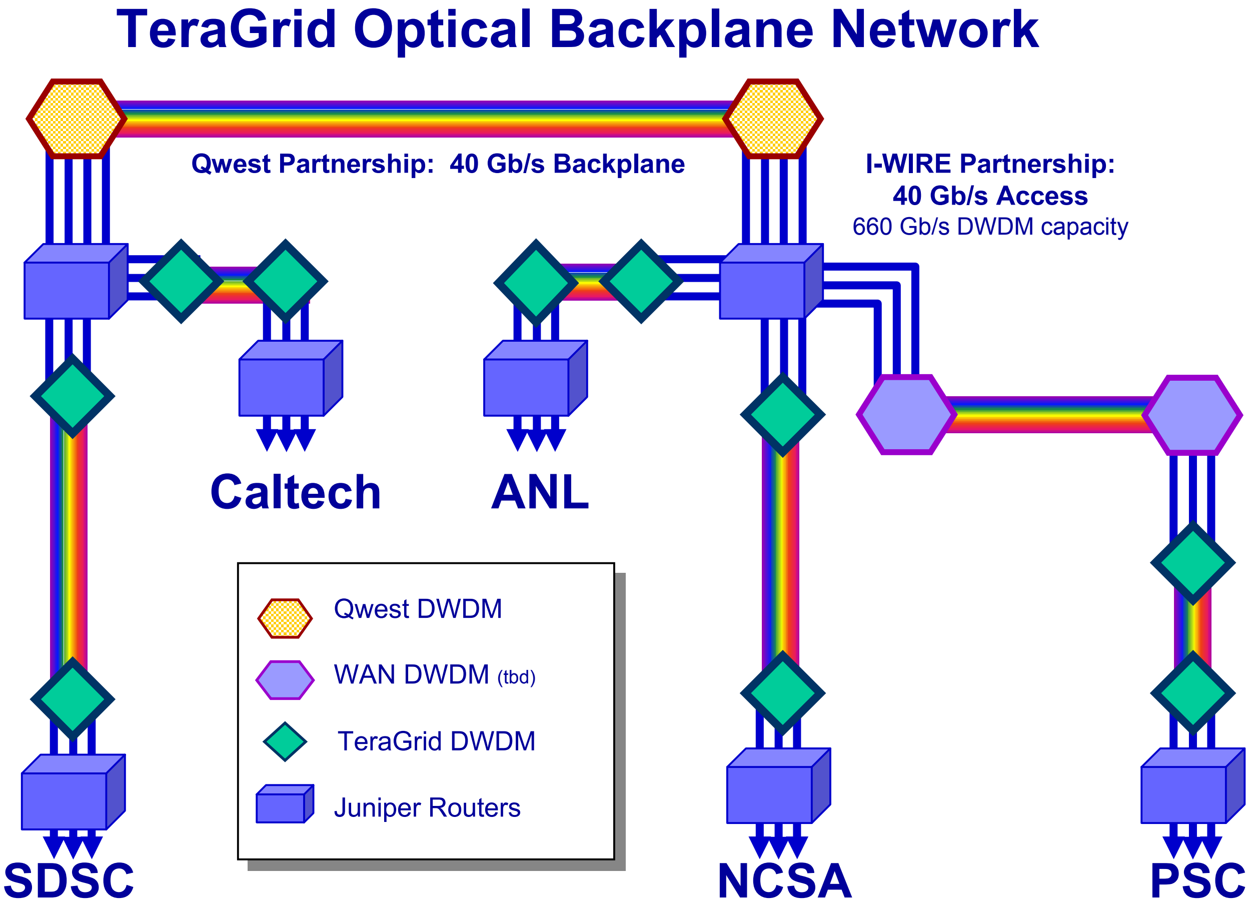
Grid Benchmarking: Why Implies How

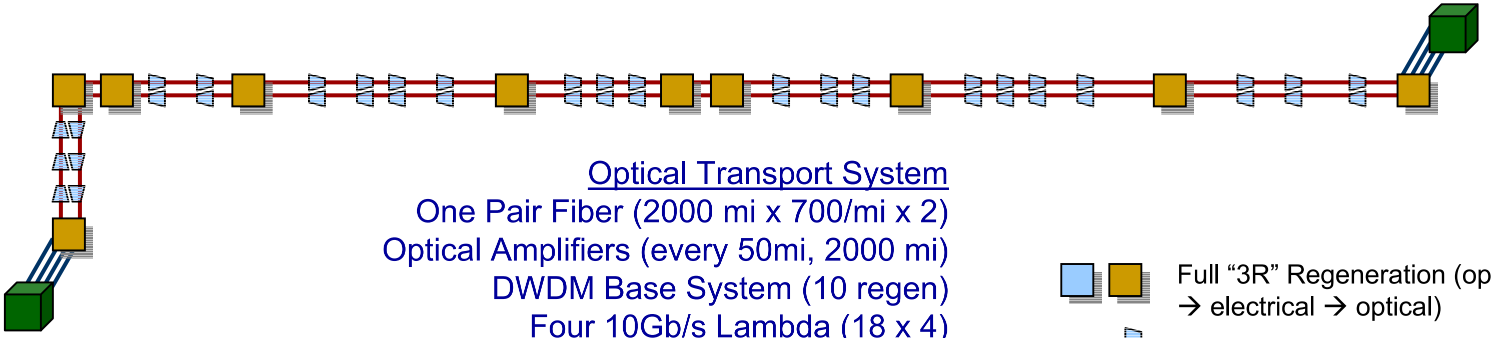
William Gropp
Mathematics and Computer Science
Argonne National Laboratory
http://www.mcs.anl.gov/~gropp

ARGONNE NATIONAL LABORATORY • UNIVERSITY OF CHICAGO •

A Simple Grid



Example 2000-mile Optical Network



Optical Transport System

One Pair Fiber (2000 mi x 700/mi x 2)

Optical Amplifiers (every 50mi, 2000 mi)

DWDM Base System (10 regen)

Four 10Gb/s Lambda (18 x 4)

Integration with LANs

IP Routers (2)

WAN interfaces (8)

LAN interfaces (8)

🟦 🟧 Full “3R” Regeneration (optical → electrical → optical)

▷ Optical amplifiers.

— LAN circuit.

— Strand of fiber

🟩 IP Router

Why Benchmark?

- Evaluate systems and approaches
- Implement individual grid-applications
- Make applications work
    - Performance requirements for realtime collaboration
    - Performance requirements to provide value over replication of resources
    - Don’t Forget Correctness/Completeness
        - Cost of security
        - Cost of added reliability layers (e.g., TCP checksums are inadequate for TB files)
        - Ability to handle grid realities (e.g., firewalls, proxies)
    - Deficiency analysis — identify large gaps between achieved and achievable performance
- Applications Requirements for Grids Should Guide Benchmarks
    - Quantify Bandwidth
    - Quantify Usability

What’s Different About the Grid?

- Shared resource
  - No reproducibility of experimental conditions
    - Classic MPPs have very good reproducibility
  - Network Weather Service — What more needs to be said
- Very complex paths for messaging; multiple transport types
- Very high latency
  - Leads to asynchronous applications
  - Performance goals emphasize bulk or realtime performance
- Often a greater software gap between the application and the hardware

Three Cautionary Tales

- LINPACK
  - Over-emphasized raw flop rate on algorithm with n^3 work on n^2 data
- SPEC (and vendor chosen tests)
  - Used to design tomorrow’s hardware for yesterday’s algorithms
- Latency/bandwidth for message-passing
  - Latency $\approx \lim_{n\to 0}$ time for message of length n, not the 0-byte time
  - Combines latency and overhead
  - Ignores contention and resource limits

Analytic Models

- Key to providing a framework for benchmarking, but
  - ♦ Must be relate to applications
- Define
  - ♦ Usefulness $= \dfrac{\text{lower bound}}{\text{upper bound}}$ $\qquad \dfrac{-1}{\ln\dfrac{\text{lower bound}}{\text{upper bound}}}$
  - ♦ Predictability $= \dfrac{\text{observed lower bound}}{\text{observed upper bound}}$
- The challenge is to be useful for systems with poor predictability while retaining simplicity
  - ♦ While you're at it, I'd like FTL and Immortality ☺

Which Benchmarks?

- What are the important application classes?
  - Data sharing
    - Benchmark file transfers with realistic sizes, security, reliability guarantees
  - Computational resource sharing
    - Workload processed; include all data staging
  - Collaboration
  - Others

Grid Challenges

- End-to-End Benchmarks
  - Strong effect from "last meter" (poor I/O to disk; saturated memory system; misconfigured interior network)
  - Must isolate effects
    - Lets corrections can be made
- Lack of Reproducibility
  - Benchmarks on the grid are experiments in the field
    - Impossible to control all factors
    - Experiments must have a valid *statistical* design
    - No "instant gratification" benchmarks
- Lack of Established Applications
  - What should we measure?

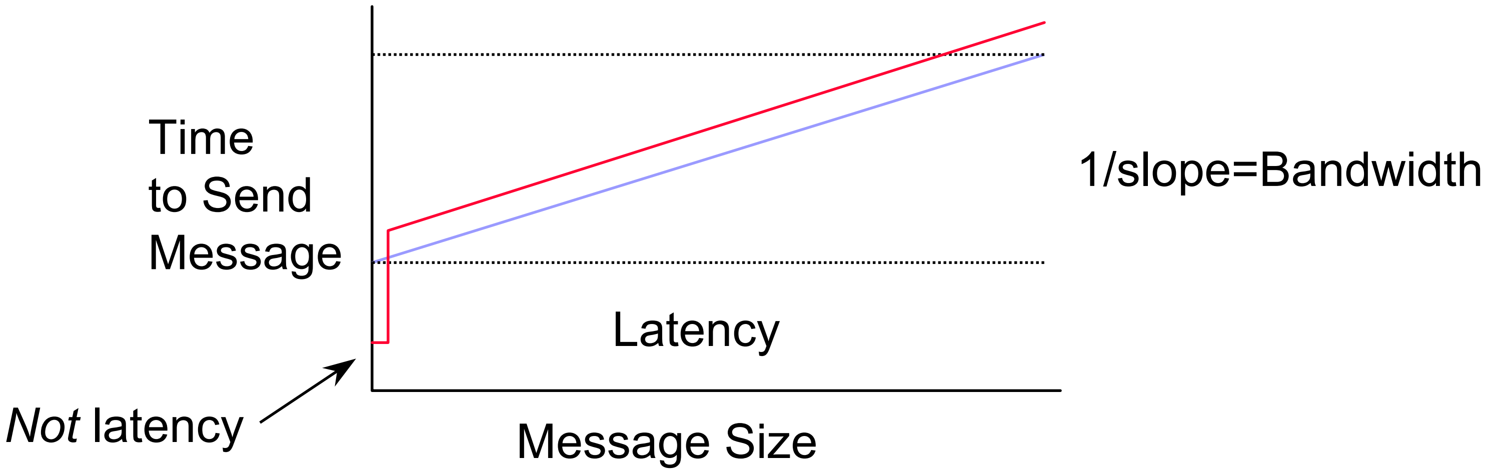
Thoughts
(less than recommendations)

- Determine critical application classes and how performance impacts their success.
    - ♦ Derive benchmark needs from these
    - ♦ Measurements and predictions must include uncertainty
- Grid simulations are needed for reproducible, controlled experiments
    - ♦ Understand effects and provides a way to evaluate new methods
    - ♦ Counterpart to lab experiments
- "Live" Grid performance measurements based on good experimental design
    - ♦ Will be statistical
    - ♦ (CS curriculum needs a course in statistics)
- Include measures of tool usability
    - ♦ Anyone remember Veronica? Gopher?

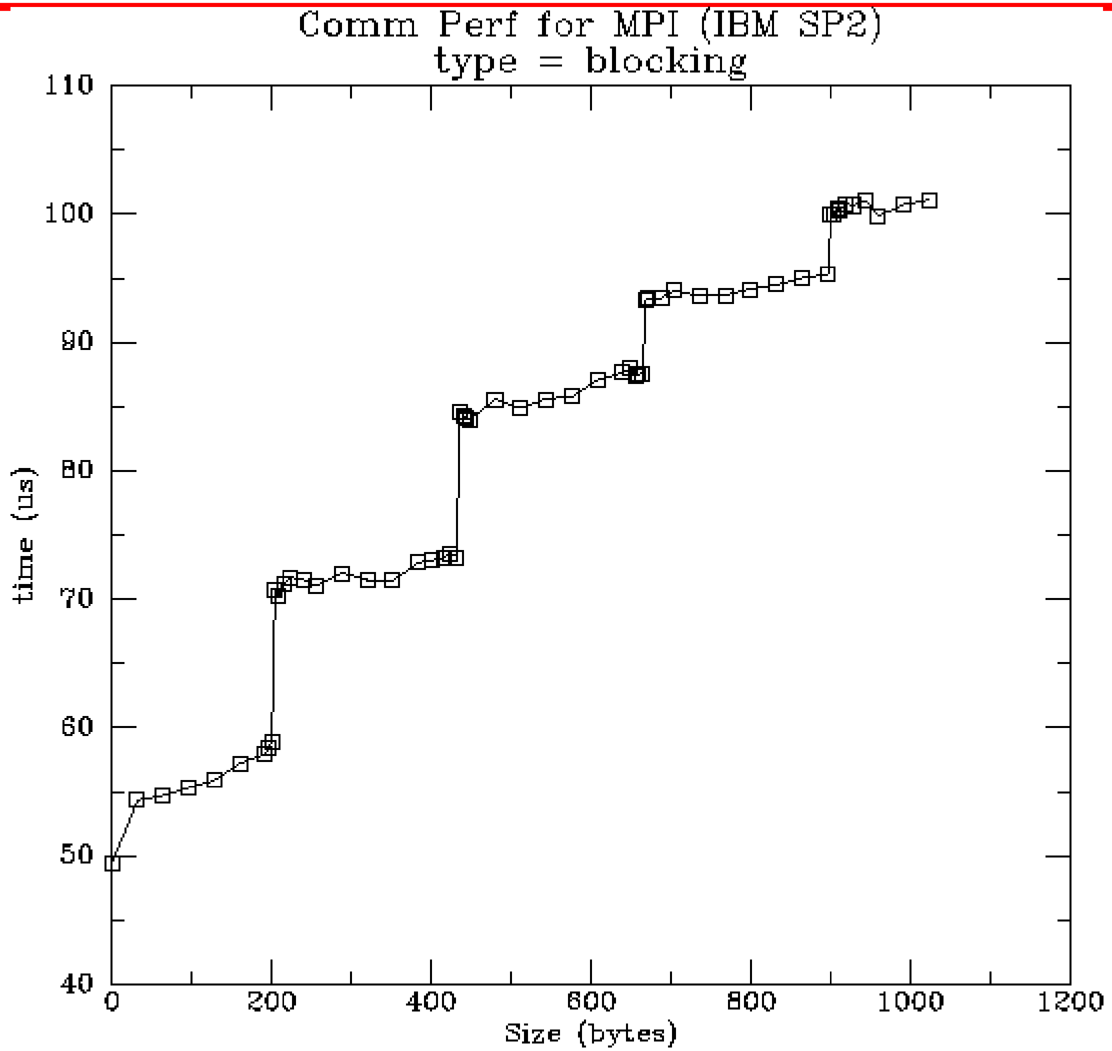
Interpreting Latency and Bandwidth

- Bandwidth is the inverse of the slope of the line

  time = latency + (1/rate) size_of_message
- For performance estimation purposes, latency is the limit(n➲0) of the time to send n bytes
- Latency is sometimes described as “time to send a message of zero bytes”. This is true *only* for the simple model. The number quoted is sometimes misleading.



Example of Packetization



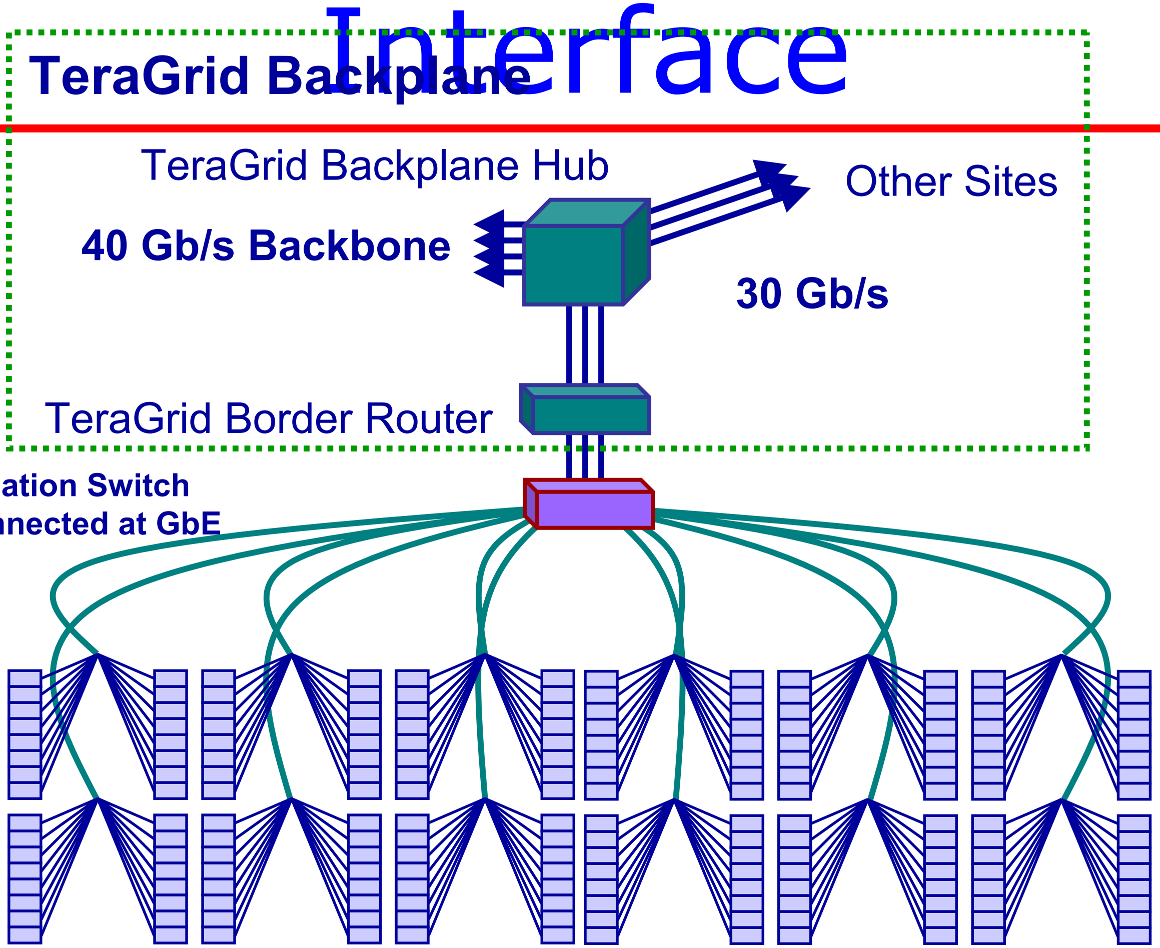
Packets contain 232 bytes of data. (first is 200 bytes, so MPI header is probably 32 bytes).

Data from mpptest, available at ftp://ftp.mcs.anl.gov/pub/mpi/misc/perftest.tar.gz

Misc Issues

- Reproducibility
  - Reproduce the grid ?!*#$
    - Testbeds : *simulate* the grid
  - Based on testbed results, conduct experiments in the field
    - *Analyze* the results; make the experiments statistically valid
- Representation
  - Choose the right benchmarks
  - Shared network
  - Are endpoints important?
- Guidance
  - Does the benchmark indicate what needs to be done to improve performance?
  - How can you react to the benchmark

DTF Cluster-Backplane Interface



TeraGrid Backplane

TeraGrid Backplane Hub

Other Sites

40 Gb/s Backbone

30 Gb/s

TeraGrid Border Router

GbE/10GbE Aggregation Switch
All nodes directly connected at GbE

ETF: PSC TCS1 will employ multiple dedicated I/O gateway nodes between internal Quadrics switch and GbE for Backplane I/O

Intra-Cluster I/O via Myrinet (not shown)

Choosing A Benchmark

- Measurements must match some model of application performance or goals